Origin of modern syphilis and emergence of a pandemic Treponema pallidum cluster

Arora, Natasha ; Schuenemann, Verena J ; Jäger, Günter ; Peltzer, Alexander ; Seitz, Alexander ; Herbig, Alexander ; Strouhal, Michal ; Grillová, Linda ; Sánchez-Busó, Leonor ; Kühnert, Denise ; Bos, Kirsten I ; Davis, Leyla Rivero ; Mikalová, Lenka ; Bruisten, Sylvia ; Komericki, Peter ; French, Patrick ; Grant, Paul R ; Pando, María A ; Vaulet, Lucía Gallo ; Fermepin, Marcelo Rodríguez ; Martinez, Antonio ; Centurion Lara, Arturo ; Giacani, Lorenzo ; Norris, Steven J ; Šmajs, David ; Bosshard, Philipp P ; González-Candelas, Fernando ; Nieselt, Kay ; Krause, Johannes ; Bagheri, Homayoun C

DOI: https://doi.org/10.1038/nmicrobiol.2016.245

Origin of modern syphilis and emergence of a pandemic Treponema pallidum cluster

Natasha Arora[1,2*], Verena J. Schuenemann[3], Günter Jäger[4], Alexander Peltzer[3,4], Alexander Seitz[4], Alexander Herbig[4], Michal Strouhal[5], Linda Grillová[5], Leonor Sánchez-Busó[6], Denise Kühnert[7], Kirsten I. Bos[3], Nada Juricevic[8], Leyla Rivero Davis[1], Lenka Paštěková [5], Sylvia Bruisten[9], Peter Komericki[10], Patrick French[11], Paul Grant[12], María A. Pando[13], Lucía Gallo Vaulet[14], Marcelo Rodríguez Fermepin[14], Antonio Martinez[15], Arturo Centurion Lara[16], Lorenzo Giacani[16], Steven J. Norris[17], David Smâjs[5], Philipp P. Bosshard[8], Fernando González-Candelas[6]*, Kay Nieselt[4]*, Johannes Krause[3]* and Homayoun C. Bagheri[1]*

Affiliations:

[1]Institute for Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland

[2]Zurich Institute of Forensic Medicine, University of Zurich, Zurich, Switzerland

[3]Institute for Archaeological Sciences, University of Tübingen, 72070 Tübingen,Germany

[4]Center for Bioinformatics, University of Tübingen, Tübingen, Germany

[5]Department of Biology, Faculty of Medicine, Masaryk University, Brno, Czech Republic

[6]Unidad Mixta Infección y Salud Pública FISABIO/Universidad de Valencia. CIBER in Epidemiology and Public Health, Spain.

[7]Department of Biosystems Science and Engineering, Computational Evolution, ETH Zurich

[8]Department of Dermatology, University Hospital of Zurich, Zurich, Switzerland

[9]Public Health Laboratory, GGD Amsterdam, Cluster Infectious Diseases, Amsterdam, the Netherlands

[10]Department of Dermatology, Medical University of Graz, Graz, Austria

[11]The Mortimer Market Centre CNWL, Camden Provider Services

[12]Department of Clinical Microbiology and Virology, University College London Hospitals NHS Foundation Trust, London

[13]Centro Nacional de Referencia para el SIDA, Departamento de Microbiología, Parasitología e Inmunología, Facultad de Medicina, Universidad de Buenos Aires, Buenos Aires, Argentina

[14]Inmunología Clínica, Departamento de Bioquímica Clínica, Facultad de Farmacia y Bioquímica, Universidad de Buenos Aires, Buenos Aires, Argentina.

[15]Clinica Dermatologica Aliaga, Valencia, Spain

[16]University of Washington, Department of Medicine, Division of Allergy and Infectious Diseases, and Department of Global Health, Seattle (WA), USA

[17]Department of Pathology and Laboratory Medicine, UTHealth McGovern Medical School, Houston, TX USA

*Corresponding authors. Email: natasha.arora@uzh.ch (N.A.); fernando.gonzalez@uv.es, kay.nieselt@uni-tuebingen.de (K.N.); johannes.krause@uni-tuebingen.de (J.K.); homayoun.bagheri@outlook.com (H.C.B.).

Abstract: Syphilis swept across the world in the 16th century as one of most prominent documented pandemics and is re-emerging worldwide despite the availability of effective antibiotics. Little is known about the genetic patterns in current infections or the evolutionary origins of the disease due to the non-cultivable and clonal nature of the causative bacterium *Treponema pallidum* subsp. *pallidum*. In this study, we used DNA capture and next generation sequencing to obtain whole genome data from syphilis patient specimens and from treponemes propagated in the lab. Phylogenetic analyses indicate that the syphilis strains examined share a common ancestor posterior to the 15th century. Moreover, most contemporary strains are azithromycin resistant, and form part of a global dominant cluster that began diversifying from a common ancestor only in the mid-20th century. This cluster has the population genetic and epidemiological features indicative of the emergence of a pandemic lineage.

One Sentence Summary: Direct genomic sequencing of treponemal DNA from syphilis patients and from laboratory strains reveals a post-Columbian ancestor for all syphilis samples, and the contemporary prevalence of globally successful lineage.

Main Text:

The abrupt onslaught of the syphilis pandemic starting in the late 15th century consolidated this devastating infectious disease as one of the most feared in human history. The first reported outbreaks in Europe were during the War of Naples in 1495 (*1*). Subsequently the epidemic spread to other continents turning into a severe health burden until the discovery of penicillin in the 20th century. This bacterial infection causes system damage to the body through the dissemination of the agent *Treponema pallidum* subsp. *pallidum* (TPA). Surprisingly, the last few decades have witnessed a dramatic global re-emergence, with annual incidence reaching 10.6 million (*2*). This resurgence despite the availability of antibiotics is striking in high-income western nations such as Switzerland, the UK, and the USA (*3*, *4*). Furthermore, while resistance to penicillin has not been identified, there has been an increase in strains not responding to the second line antibiotic azithromycin (*4*).

The epidemiological characteristics of this re-emergence are poorly understood, particularly the underlying patterns of genetic diversity. Obtaining genetic data for TPA is hindered by the low quantities of endogenous DNA in clinical samples, and inability to cultivate the pathogen (*5*). Consequently, much of our molecular understanding comes from propagating strains in laboratory animals to obtain sufficient DNA. The few published whole genomes were obtained after amplification through rabbit passage (*6*–*8*), and represent limited diversity for phylogenetic analyses. These sequences indicate that the TPA genome of 1.14 Mb is genetically monomorphic. The genetic diversity, while low, remains unexplored because clinical samples are mostly typed by PCR amplification of only 1-5 loci from roughly 1000 genes in the genome (*9*, *10*). These genetic studies are mainly epidemiological and driven by the inability of serologic or microscopic tests to distinguish among TPA strains or among the subspecies *Treponema pallidum* subsp. *pertenue* (TPE) and *Treponema pallidum* subsp. *endemicum* (TEN), which cause the diseases yaws and bejel, respectively. All three diseases are transmitted through skin contact and show an overlap in their clinical manifestations. While syphilis is geographically widespread and generally transmitted sexually, yaws and bejel are mainly found endemically in hot climates and primarily transmitted between children by incidental skin contact (*11*). The precise relationships among the bacteria are still debated, particularly the evolutionary origin of syphilis.

The paucity of molecular studies and the focus on epidemiological typing of a few genes means that we are often in the dark as to the evolution and spread of epidemic TPA across the globe. To assess genomic variation in syphilis infections, we utilized DNA capture techniques (*12*) to enrich for treponemal DNA before high throughput sequencing. In total, we obtained 70 samples from 13 countries, including 52 syphilis swabs collected directly from patients between 2012 and 2013, and 18 syphilis, yaws, and bejel samples collected from 1912 onwards and propagated in laboratory rabbits (Table S1). By examining whole genome variation and reconstructing phylogenies, our results shed light on the evolutionary history of TPA and identify epidemiologically relevant haplotypes.

The distinct evolutionary histories of treponemal lineages

Due to the large background of host DNA, extracts from all 70 samples were turned into Illumina sequencing libraries and enriched genome-wide using DNA array hybridization capture prior to sequencing (*12*). The obtained 483,450 to 100,414,614 reads were mapped to the TPA reference genome (Nichols, NC_021490; Table S2). Genomic coverage ranged from 0.13-fold to over 1000-fold, with the highest genome coverage in laboratory strains propagated in rabbits, and highest variation in the samples collected from patients (0.13-fold to 223-fold) (Table S3 and Supplementary materials). To ensure robust inferences in genome-wide analyses, we incorporated the 28 samples where at least 80% of the genome was covered by a minimum of three reads (supplementary text). We investigated structural variation by using the sequencing reads of the four highest covered syphilis swab samples (NE17, NE20, CZ27, AU15) and one Indonesian yaws isolate (IND1) for a *de novo* genome assembly. Gaps were expected for the 8% of the genome containing repetitive regions and related genes such as the *tpr* subfamilies or the ribosomal RNA operons. We found no significant structural changes in the five genomes (Fig.1A), except for the deletion in IND1 of the virulence-factor encoding candidate gene TP1030. The deletion was shared across all the yaws infection isolates (supplementary text), consistent with other studies (*13*).

Prior to phylogenetic reconstruction, we checked for signatures of non-vertical descent. While *T. pallidum* is considered to be a clonal species (*14*), previous studies suggest recombinant genes in a Mexican syphilis and a Bosnian bejel isolate (*8*, *15*). We examined each of the 978 annotated genes across a total of 39 genomes, adding to our 28 sequenced genomes the 11 publicly available genomes from laboratory strains (Table S2). Genes were considered putative recombinants if they satisfied 3 conditions: i) had twice the expected SNP density as compared to the average distribution, ii) produced gene tree topologies incongruent with that of the genome-wide tree, and iii) had four or more homoplasies in at least 2 branches. This procedure identified 4 genes coding for outer membrane protein genes (Table S6) two of which (TP0136 and TP0326) are used in typing studies (*6*).

The independent history and origin of TPA

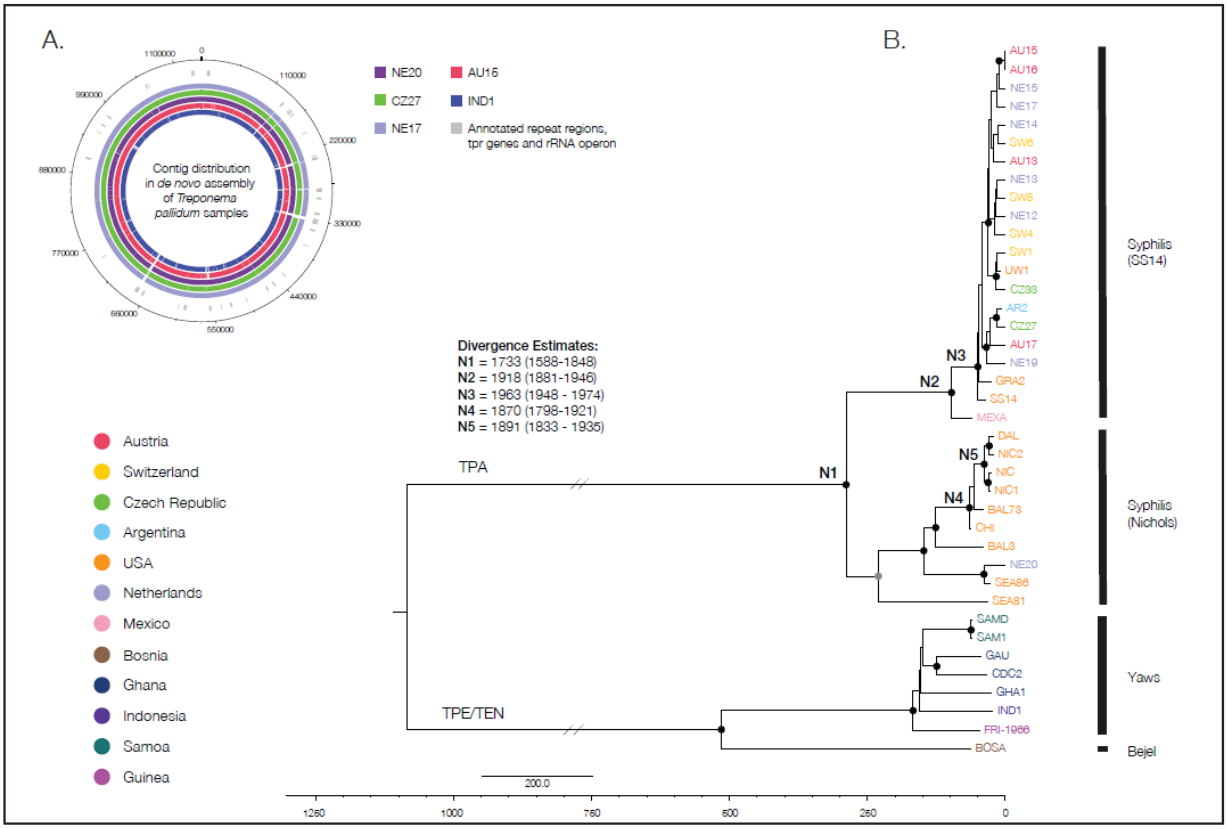
After removing the 4 putative recombinant genes from the genome alignment of all 39 samples, we reconstructed the phylogeny using the Bayesian framework implemented in BEAST (*16*). As illustrated in Fig. 1B, the phylogenetic tree revealed a marked separation of TPA from TPE/TEN, with TPA forming a monophyletic lineage. The distinction of the two lineages was robust even with the inclusion of the putative recombinant genes in the tree inference (Fig. S2). Analyses of divergence between the two lineages yielded an average mean distance of 1225 mutations. By contrast, within each of the lineages

we found at least 5 times less divergence (124.6 average pairwise mutations within the TPA lineage and 200.2 within TPE/TEN). These results highlight the perils of relying on a limited set of markers for taxonomic classification or typing schemes, which may yield spurious groupings between TPA and TPE as observed for the TP0548 typing gene (Fig. S2).

Using the isolation dates for the TPA samples as tip calibration and applying the Birth Death Serial Skyline model (*17*), we obtained a scaled mean evolutionary rate of 6.6×10^{-7} substitutions per site per year for the whole genome, in line with estimates for other clonal human pathogens such as *Shigella sonnei* (6.0×10^{-7}) and *Vibrio cholerae* (01 lineages; 8.0×10^{-7}) (*18*, *19*).

Our divergence analyses for TPA samples provide a time to the most recent common ancestor (TMRCA) less than 500 years ago (mean calendar year 1733, 95% HPD 1588-1848; Fig. 1B), no earlier than the syphilis pandemic starting in the late 15th century. While our analyses do not exclude the possibility that older TPA strains existed in Europe prior to the pandemic we detect no evidence for it. Importantly, a lasting genetic signature was left by a syphilis ancestor that existed during the Renaissance pandemic.

Figure 1. *De novo* genome assemblies and phylogenetic reconstruction. A) *De novo* genome assembly for four syphilis patient and one yaws sample, with color coded geographic origin (inset legend). Blank spaces correspond to gaps, overlapping with gene regions that are difficult to assemble from short reads such as the *tpr* subfamilies and rRNA operons (regions shown in the outermost ring in gray). B) BEAST tree for the 39 genomes (after excluding putatively recombinant genes), with black circles indicating nodes with ≥99% posterior probabilities (PP); gray circle denotes 82% PP. Divergence date estimates (median and 95% highest posterior density) for major well-supported TPA nodes are given in the legend. What is the axes?



Rapid spread of a contemporary epidemic cluster

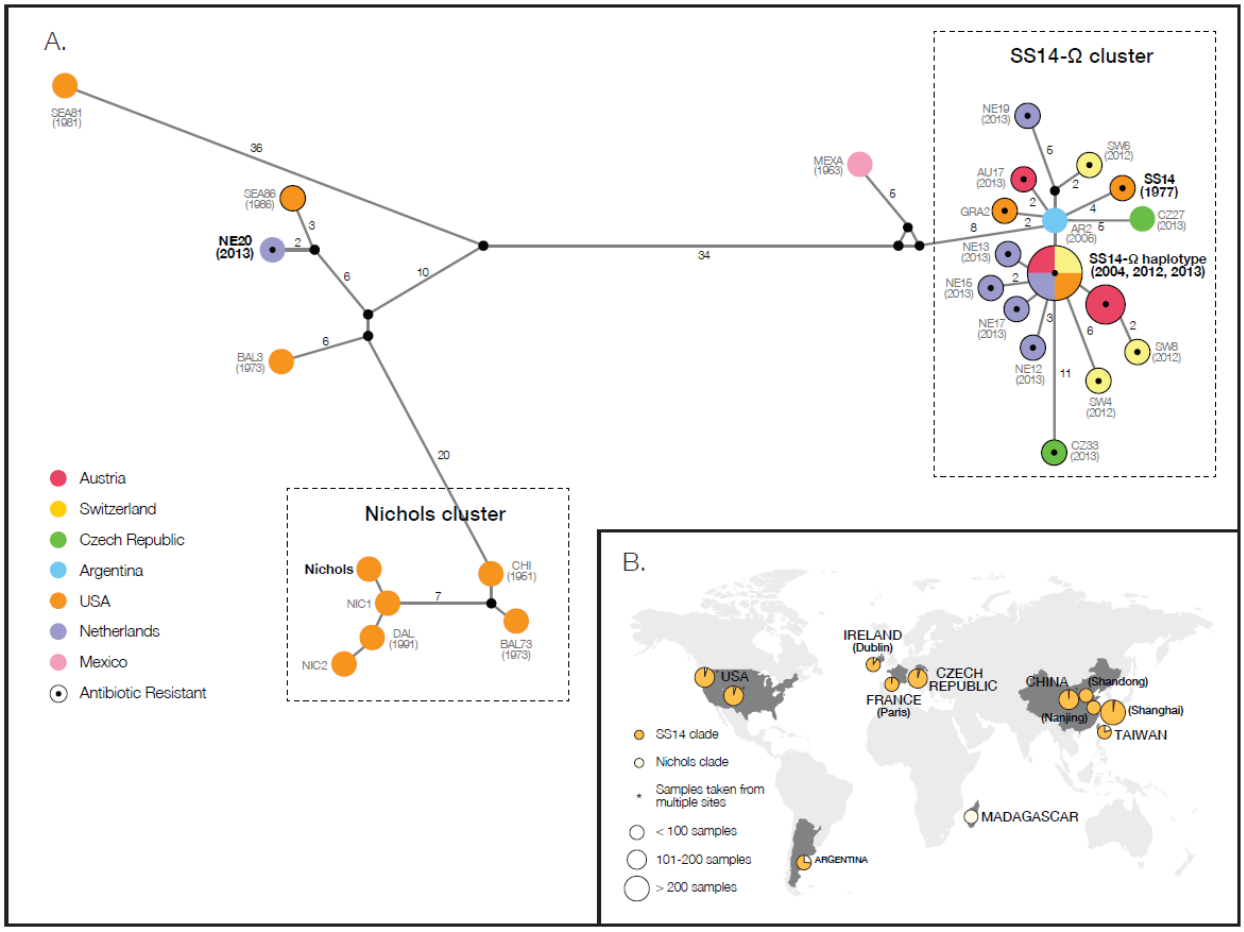
A median-joining (MJ) network for the 31 TPA samples highlights the mutational differences between the Nichols and the SS14 clades, named after reference genomes, and distinguishes several clusters (Fig. 2A, Fig. 1B). The Nichols clade consists almost exclusively of samples collected from patients in North America from 1912 to 1986, passaged in rabbits prior to sequencing, with the exception of one patient sample from 2013. In contrast, the SS14 clade has a more global distribution, encompassing European, North American and South American samples collected from infections between 1951 and 2013. Strikingly, the SS14 clade contains a dominant central haplotype (labelled as SS14-Ω in Fig. 2A) from which the other sequences radiate. The cluster associated with the SS14- Ω haplotype contains all but one of the recent patient samples from 2012-2013 (n=17/18) that were captured and sequenced directly, in addition to three samples from 1977 (n=1) and 2004 (n=2). Our dating analyses point to an origin for the SS14- Ω cluster in the second half of the 20th century (median calendar year 1963, 95% HPD 1948-1974), at a time when incidence was reduced due to the introduction of antibiotics.

To check whether the prevalence of SS14 clade sequences applies to countries not represented by our data, we examined sequences for the widely typed TP0548 hypervariable gene region in epidemiological

typing studies (*9*). This gene distinguishes the SS14 from the Nichols clade among TPA samples (Supplementary Section 7; Fig. S5?). Across 970 publicly available global sequences (Supplementary Section 7) we found that 93.5% of them fell into the SS14 clade, consistent with our findings on the recent spread of an epidemic cluster. The wide geographical presence of the SS14 clade establishes it as representative of the present worldwide epidemic. While studies to date have focused on the Nichols strain (*20*, *21*), our results indicate that further focus on the SS14 clade is warranted.

Typing of samples collected across several different years in the Czech Republic, San Francisco, British Columbia and Seattle indicate that macrolide antibiotic resistance has increased over time (*4*, *10*, *22*–*24*). We queried the presence of the two mutations (A2058G and A2059G) in the 23S rRNA genes associated with azithromycin resistance (*4*, *25*). As observed in the MJ network, the resistance marker is a dominant characteristic of the SS14- Ω cluster (Fig 2B), although it is also found in the recent Dutch sample of the Nichols clade. Extending our analyses of the 23S rRNA gene to all sequenced samples from our study, including those with low coverage, revealed the mutations in 90% of the SS14 and 25% of the Nichols samples, indicating that neither resistance nor sensitivity is clade-specific. Hence resistance was not an ancestral characteristic of the SS14 clade. A likely explanation is the extensive usage of azithromycin to treat syphilis and a wide range of bacterial infections, including co-infections with other sexually-transmitted diseases, which can play an important role in the selection and subsequent spread of resistance (*26*).

Figure 2. **Median-joining (MJ) network analysis and geographic distribution of the SS14 and Nichols clades.** A) Median-joining network for genome-wide variable positions, excluding sites with missing data (n=628). Circles represent haplotypes, with geographical origin color-coded. Number of mutations, when above one, is given on the branches. Inferred haplotypes (median vectors) are shown as black connecting circles. Central black circles within haplotypes indicate azithromycin resistance marker. B) Relative frequencies of SS14 clade versus Nichols clade isolates typed across global studies is shown in pie charts, with sizes proportional to sample size.



Conclusions

We examined the genomic diversity of *T. pallidum* subsp. *pallidum* (TPA) from syphilis samples isolated during the 20th and 21st centuries. The results include the first reported whole genome sequences successfully obtained directly from syphilis patients. Our analyses indicate that all TPA samples examined to date share a common ancestor that was infecting populations within the early centuries of the modern era. The present work does not necessarily resolve the question whether ancestral TPA originated in the Americas or Europe. However, our results suggest that the events posterior to the colonization of the Americas provided the context for the spread of an STD lineage that acquired pandemic proportions. Furthermore, our analyses have identified a present-day epidemic cluster (SS14-Ω) that displays the population genetic and epidemiological features of an emergent pandemic.

References and Notes

1. C. Quétel, *History of syphilis* (Johns Hopkins University Press, Baltimore, 1990).

2. J. Rowley *et al.*, *Global incidence and prevalence of selected curable sexually transmitted infections, 2008* (World Health Organization, Geneva, Switzerland, 2012; http://apps.who.int/iris/bitstream/10665/75181/1/9789241503839_eng.pdf).

3. K. A. Fenton *et al.*, Infectious syphilis in high-income settings in the 21st century. *Lancet Infect. Dis.* **8**, 244–253 (2008).

4. L. V. Stamm, Global Challenge of Antibiotic-Resistant Treponema pallidum. *Antimicrob. Agents Chemother.* **54**, 583–589 (2010).

5. C. M. Fraser *et al.*, Complete genome sequence of Treponema pallidum, the syphilis spirochete. *Science*. **281**, 375–88 (1998).

6. D. Šmajs, S. J. Norris, G. M. Weinstock, Genetic diversity in Treponema pallidum: Implications for pathogenesis, evolution and molecular diagnostics of syphilis and yaws. *Infect. Genet. Evol.* **12**, 191–202 (2012).

7. L. Giacani *et al.*, Complete Genome Sequence of the Treponema pallidum subsp. pallidum Sea81-4 Strain. *Genome Announc.* **2**, e00333–14–e00333–14 (2014).

8. B. Štaudová *et al.*, Whole Genome Sequence of the Treponema pallidum subsp. endemicum Strain Bosnia A: The Genome Is Related to Yaws Treponemes but Contains Few Loci Similar to Syphilis Treponemes. *PLoS Negl. Trop. Dis.* **8**, e3261 (2014).

9. C. M. Marra *et al.*, Enhanced Molecular Typing of Treponema pallidum: Geographical Distribution of Strain Types and Association with Neurosyphilis. *J. Infect. Dis.* **202**, 1380–1388 (2010).

10. L. Grillová *et al.*, Molecular Typing of Treponema pallidum in the Czech Republic during 2011 to 2013: Increased Prevalence of Identified Genotypes and of Isolates with Macrolide Resistance. *J. Clin. Microbiol.* **52**, 3693–3700 (2014).

11. L. Giacani, S. A. Lukehart, The Endemic Treponematoses. *Clin. Microbiol. Rev.* **27**, 89–115 (2014).

12. E. Hodges *et al.*, Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. *Nat. Protoc.* **4**, 960–974 (2009).

13. L. Mikalova *et al.*, Genome analysis of Treponema pallidum subsp. pallidum and subsp. pertenue strains: most of the genetic differences are localized in six regions. *PLoS One*. **5**, e15713 (2010).

14. M. Achtman, Evolution, Population Structure, and Phylogeography of Genetically Monomorphic Bacterial Pathogens. *Annu. Rev. Microbiol.* **62**, 53–70 (2008).

15. H. Petrosova *et al.*, Whole genome sequence of Treponema pallidum ssp. pallidum, strain Mexico A, suggests recombination between yaws and syphilis strains. *PLoS Negl Trop Dis*. **6**, e1832 (2012).

16. R. Bouckaert *et al.*, BEAST 2: A Software Platform for Bayesian Evolutionary Analysis. *PLoS Comput. Biol.* **10**, e1003537 (2014).

17. T. Stadler, D. Kuhnert, S. Bonhoeffer, A. J. Drummond, Birth-death skyline plot reveals temporal changes of epidemic spread in HIV and hepatitis C virus (HCV). *Proc. Natl. Acad. Sci.* **110**, 228–233 (2013).

18. K. E. Holt *et al.*, Shigella sonnei genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. *Nat Genet*. **44**, 1056–1059 (2012).

19. A. Mutreja *et al.*, Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature*. **477**, 462–465 (2011).

20. L. Giacani *et al.*, Footprint of Positive Selection in Treponema pallidum subsp. pallidum Genome Sequences Suggests Adaptive Microevolution of the Syphilis Pathogen. *PLoS Negl. Trop. Dis.* **6**, e1698 (2012).

21. M. Strouhal *et al.*, Genome Differences between Treponema pallidum subsp. pallidum Strain Nichols and T. paraluiscuniculi Strain Cuniculi A. *Infect. Immun.* **75**, 5859–5866 (2007).

22. C. M. Marra *et al.*, Antibiotic selection may contribute to increases in macrolide-resistant Treponema pallidum. *J. Infect. Dis.* **194**, 1771–1773 (2006).

23. S. J. Mitchell *et al.*, Azithromycin-resistant syphilis infection: San Francisco, California, 2000–2004. *Clin. Infect. Dis.* **42**, 337–345 (2006).

24. M. Morshed, H. Jones, Treponema pallidum macrolide resistance in BC. *CMAJ*, 349 (2006).

25. P. Matejkova *et al.*, Macrolide treatment failure in a case of secondary syphilis: a novel A2059G mutation in the 23S rRNA gene of Treponema pallidum subsp. pallidum. *J. Med. Microbiol.* **58**, 832–836 (2009).

26. D. Šmajs, L. Paštěková, L. Grillová, Macrolide Resistance in the Syphilis Spirochete, Treponema pallidum ssp. pallidum: Can We Also Expect Macrolide-Resistant Yaws Strains? *Am. J. Trop. Med. Hyg.* **93**, 678–683 (2015).

Acknowledgments

Research in Zurich was funded by the *Forschungskredit* and the University of Zurich. We thank Alexei Drummond for input on BEAST. A.H was funded by the European Research Council Starting Grant. Work at the University of Valencia was funded by Ministerio de Economía y Competitividad. Special thanks to Lukas Keller and his lab for their support. We are also grateful to Stephan Lautenschlager for his initial guidance.